CHAPTER 5
Cosmids, phasmids and other advanced vectors

**Introduction**

In the 1970s, when recombinant DNA technology was first being developed, only a limited number of vectors were available and these were based on either high-copy-number plasmids or phage \( \lambda \). Later, phage M13 was developed as a specialist vector to facilitate DNA sequencing. Gradually, a number of purpose-built vectors were developed, of which pBR322 is probably the best example, but the creation of each one was a major task. Over time, a series of specialist vectors was constructed, each for a particular purpose. During this period, there were many arguments about the relative benefits of plasmid and phage vectors. Today, the molecular biologist has available an enormous range of vectors and these are notable for three reasons. First, many of them combine elements from both plasmids and phages and are known as phasmids or, if they contain an M13 ori region, phagemids. Secondly, many different features that facilitate cloning and expression can be found combined in a single vector. Thirdly, purified vector DNA plus associated reagents can be purchased from molecular-biology suppliers. The hapless scientist who opens a molecular-biology catalogue is faced with a bewildering selection of vectors and each vender promotes different ones. Although the benefits of using each vector may be clear, the disadvantages are seldom obvious. The aim of this chapter is to provide the reader with a detailed explanation of the biological basis for the different designs of vector.

There are two general uses for cloning vectors: cloning large pieces of DNA and manipulating genes. When mapping and sequencing genomes, the first step is to subdivide the genome into manageable pieces. The larger these pieces, the easier it is to construct the final picture (see Chapter 7); hence the need to clone large fragments of DNA. Large fragments are also needed if it is necessary to ‘walk’ along the genome to isolate a gene, and this topic is covered in Chapter 6. In many instances, the desired gene will be relatively easy to isolate and a simpler cloning vector can be used. Once isolated, the cloned gene may be expressed as a probe sequence or as a protein, it may be sequenced or it may be mutated \( \textit{in vitro} \). For all these applications, small specialist vectors are used.

**Vectors for cloning large fragments of DNA**

**Cosmid vectors**

As we have seen, concatemers of unit-length \( \lambda \) DNA molecules can be efficiently packaged if the cos sites, substrates for the packaging-dependent cleavage, are 37–52 kb apart (75–105% the size of \( \lambda^+ \) DNA). In fact, only a small region in the proximity of the cos site is required for recognition by the packaging system (Hohn 1975).

Plasmids have been constructed which contain a fragment of \( \lambda \) DNA including the cos site (Collins & Brüning 1978, Collins & Hohn 1979, Wahl et al. 1987, Evans et al. 1989). These plasmids have been termed cosmids and can be used as gene-cloning vectors in conjunction with the \( \textit{in vitro} \) packaging system. Figure 5.1 shows a gene-cloning scheme employing a cosmid. Packaging the cosmid recombinants into phage coats imposes a desirable selection upon their size. With a cosmid vector of 5 kb, we demand the insertion of 32–47 kb of foreign DNA – much more than a phage-\( \lambda \) vector can accommodate. Note that, after packaging \( \textit{in vitro} \), the particle is used to infect a suitable host. The recombinant cosmid DNA is injected and circularizes like phage DNA but replicates as a normal plasmid without the expression of any phage functions. Transformed cells are selected on the basis of a vector drug-resistance marker.
Cosmids, phasmids and advanced vectors

Cosmids provide an efficient means of cloning large pieces of foreign DNA. Because of their capacity for large fragments of DNA, cosmids are particularly attractive vectors for constructing libraries of eukaryotic genome fragments. Partial digestion with a restriction endonuclease provides suitably large fragments. However, there is a potential problem associated with the use of partial digests in this way. This is due to the possibility of two or more genome fragments joining together in the ligation reaction, hence creating a clone containing fragments that were not initially adjacent in the genome. This would give an incorrect picture of their chromosomal organization. The problem can be overcome by size fractionation of the partial digest.

**Fig. 5.1** Simple scheme for cloning in a cosmid vector. (See text for details.)
Even with sized foreign DNA, in practice cosmid clones may be produced that contain non-contiguous DNA fragments ligated to form a single insert. The problem can be solved by dephosphorylating the foreign DNA fragments so as to prevent their ligation together. This method is very sensitive to the exact ratio of target-to-vector DNAs (Collins & Brüning 1978) because vector-to-vector ligation can occur. Furthermore, recombinants with a duplicated vector are unstable and break down in the host by recombination, resulting in the propagation of a non-recombinant cosmid vector.

Such difficulties have been overcome in a cosmid-cloning procedure devised by Ish-Horowicz and Burke (1981). By appropriate treatment of the cosmid vector pJB8 (Fig. 5.2), left-hand and right-hand vector ends are purified which are incapable of self-ligation but which accept dephosphorylated foreign DNA. Thus the method eliminates the need to size the foreign DNA fragments and prevents formation...
Cosmids, phasmids and advanced vectors

Fig. 5.3 Cosmid cloning scheme of Bates and Swift (1983). The cosmid c2XB contains two cos sites, separated by a site for the restriction endonuclease SmaI which creates blunt ends. These blunt ends ligate only very inefficiently under the conditions used and effectively prevent the formation of recombinants containing multiple copies of the vector.

An alternative solution to these problems has been devised by Bates and Swift (1983) who have constructed cosmid c2XB. This cosmid carries a BamHI insertion site and two cos sites separated by a blunt-end restriction site (Fig. 5.3). The creation of these blunt ends, which ligate only very inefficiently under the conditions used, effectively prevents vector self-ligation in the ligation reaction.

Modern cosmids of the pWE and sCos series (Wahl et al. 1987, Evans et al. 1989) contain features such as: (i) multiple cloning sites (Bates & Swift 1983, Pirrotta et al. 1983, Breter et al. 1987) for simple cloning using non-size-selected DNA; (ii) phage promoters flanking the cloning site; and (iii) unique NotI, SacII or SfiI sites (rare cutters, see Chapter 6) flanking the cloning site to permit removal of the insert from the vector as single fragments. Mammalian expression modules encoding dominant selectable markers (Chapter 10) may also be present, for gene transfer to mammalian cells if required.

BACs and PACs as alternatives to cosmids

Phage P1 is a temperate bacteriophage which has been extensively used for genetic analysis of Escherichia coli because it can mediate generalized transduction. Sternberg and co-workers have developed a P1 vector system which has a capacity for
DNA fragments as large as 100 kb (Sternberg 1990, Pierce et al. 1992). Thus the capacity is about twice that of cosmid clones but less than that of yeast artificial chromosome (YAC) clones (see p. 159). The P1 vector contains a packaging site (pac) which is necessary for in vitro packaging of recombinant molecules into phage particles. The vectors contain two loxP sites. These are the sites recognized by the phage recombinase, the product of the phage cre gene, and which lead to circularization of the packaged DNA after it has been injected into an E. coli host expressing the recombinase (Fig. 5.4). Clones are maintained in E. coli as low-copy-number plasmids by selection for a vector kanamycin-resistance marker. A high copy number can be induced by exploitation of the P1 lytic replicon (Sternberg 1990). This P1 system has been used to construct genomic libraries of mouse, human, fission yeast and Drosophila DNA (Hoheisel et al. 1993, Hartl et al. 1994).

Shizuya et al. (1992) have developed a bacterial cloning system for mapping and analysis of complex genomes. This BAC system (bacterial artificial chromosome) is based on the single-copy sex factor F of E. coli. This vector (Fig. 5.5) includes the λ cos N and P1 loxP sites. Two cloning sites (HindIII and BamHI) and several G+C restriction enzyme sites (e.g. SfiI, NotI, etc.) for potential excision of the inserts. The cloning site is also flanked by T7 and SP6 promoters for generating RNA probes. This BAC can be transformed into E. coli very efficiently, thus avoiding the
Cosmids, phasmids and advanced vectors

Packaging extracts that are required with the P1 system. BACs are capable of maintaining human and plant genomic fragments of greater than 300 kb for over 100 generations with a high degree of stability (Woo et al. 1994) and have been used to construct genome libraries with an average insert size of 125 kb (Wang et al. 1995a). Subsequently, Ioannou et al. (1994) have developed a P1-derived artificial chromosome (PAC), by combining features of both the P1 and the F-factor systems. Such PAC vectors are able to handle inserts in the 100–300 kb range.

The first BAC vector, pBAC108L, lacked a selectable marker for recombinants. Thus, clones with inserts had to be identified by colony hybridization. While this once was standard practice in gene manipulation work, today it is considered to be inconvenient! Two widely used BAC vectors, pBeloBAC11 and pECBAC1, are derivatives of pBAC108L in which the original cloning site is replaced with a lacZ gene carrying a multiple cloning site (Kim et al. 1996, Frijters et al. 1997). pBeloBAC11 has two EcoRI sites, one in the lacZ gene and one in the CMβ gene, whereas pECBAC1 has only the EcoRI site in the lacZ gene. Further improvements to BACs have been made by replacing the lacZ gene with the sacB gene (Hamilton et al. 1996). Insertional inactivation of sacB permits growth of the host cell on sucrose-containing media, i.e. positive selection for inserts. Frengen et al. (1999) have further improved these BACs by including a site for the insertion of a transposon. This enables genomic inserts to be modified after cloning in bacteria, a procedure known as retrofitting. The principal uses of retrofitting are the simplified introduction of deletions (Chatterjee & Coren 1997) and the introduction of reporter genes for use in the original host of the genomic DNA (Wang et al. 2001).

**Choice of vector**

The maximum size of insert that the different vectors will accommodate is shown in Table 5.1. The size of insert is not the only feature of importance. The absence of chimeras and deletions is even more important. In practice, some 50% of YACs show structural instability of inserts or are chimeras in which two or more DNA fragments have become incorporated into one clone. These defective YACs are unsuitable for use as mapping and sequencing reagents and a great deal of effort is required to identify them. Cosmid inserts sometimes contain

---

**Table 5.1** Maximum DNA insert possible with different cloning vectors. YACs are discussed on p. 159.

<table>
<thead>
<tr>
<th>Vector</th>
<th>Host</th>
<th>Insert size</th>
</tr>
</thead>
<tbody>
<tr>
<td>λ phage</td>
<td><em>E. coli</em></td>
<td>5–25 kb</td>
</tr>
<tr>
<td>λ cosmids</td>
<td><em>E. coli</em></td>
<td>35–45 kb</td>
</tr>
<tr>
<td>P1 phage</td>
<td><em>E. coli</em></td>
<td>70–100 kb</td>
</tr>
<tr>
<td>PACs</td>
<td><em>E. coli</em></td>
<td>100–300 kb</td>
</tr>
<tr>
<td>BACs</td>
<td><em>E. coli</em></td>
<td>≤ 300 kb</td>
</tr>
<tr>
<td>YACs</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>200–2000 kb</td>
</tr>
</tbody>
</table>
the same aberrations and the greatest problem with them arises when the DNA being cloned contains tandem arrays of repeated sequences. The problem is particularly acute when the tandem array is several times larger than the allowable size of a cosm id insert. Potential advantages of the BAC and PAC systems over YACs include lower levels of chimerism (Hartl et al. 1994, Sternberg 1994), ease of library generation and ease of manipulation and isolation of insert DNA. BAC clones seem to represent human DNA far more faithfully than their YAC or cosm id counterparts and appear to be excellent substrates for shotgun sequence analysis, resulting in accurate contiguous sequence data (Venter et al. 1996).

**Specialist-purpose vectors**

**Vectors that can be used to make single-stranded DNA for sequencing**

Whenever a new gene is cloned or a novel genetic construct is made, it is usual practice to sequence all or part of the chimeric molecule. As will be seen later (p. 120), the Sanger method of sequencing requires single-stranded DNA as the starting material. Originally, single-stranded DNA was obtained by cloning the sequence of interest in an M13 vector (see p. 60). Today, it is more usual to clone the sequence into a pUC-based phagemid vector which contains the M13 ori region as well as the pUC (Col E1) origin of replication. Such vectors normally replicate inside the cell as double-stranded molecules. Single-stranded DNA for sequencing can be produced by infecting cultures with a helper phage such as M13K07. This helper phage has the origin of replication of P15A and a kanamycin-resistant gene inserted into the M13 ori region and carries a mutation in the gII gene (Vieira & Messing 1987). M13KO7 can replicate on its own. However, in the presence of a phagemid bearing a wild-type origin of replication, single-stranded phagemid is packaged preferentially and secreted into the culture medium. DNA purified from the phagemids can be used directly for sequencing.

**Expression vectors**

Expression vectors are required if one wants to prepare RNA probes from the cloned gene or to purify large amounts of the gene product. In either case, transcription of the cloned gene is required. Although it is possible to have the cloned gene under the control of its own promoter, it is more usual to utilize a promoter specific to the vector. Such vector-carried promoters have been optimized for binding of the *E. coli* RNA polymerase and many of them can be regulated easily by changes in the growth conditions of the host cell.

*E. coli* RNA polymerase is a multi-subunit enzyme. The core enzyme consists of two identical α subunits and one each of the β and β′ subunits. The core enzyme is not active unless an additional subunit, the σ factor, is present. RNA polymerase recognizes different types of promoters depending on which type of σ factor is attached. The most common promoters are those recognized by the RNA polymerase with σ^70_. A large number of σ^70_ promoters from *E. coli* have been analysed and a compilation of over 300 of them can be found in Lisser and Margalit (1993). A comparison of these promoters has led to the formulation of a consensus sequence (Fig. 5.6). If the transcription start point is assigned the position +1 then

<table>
<thead>
<tr>
<th>-35 Region</th>
<th>-10 Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSENSUS</td>
<td>••••TTGACA•••••••••TATAT•••••••••</td>
</tr>
<tr>
<td>lac</td>
<td>GGGTTTACTTTATGCTTCCCGGCTGTATATTGT</td>
</tr>
<tr>
<td>trp</td>
<td>CTGTTGACAATTAATCAT CGAACATGTAACTAG</td>
</tr>
<tr>
<td>Xp</td>
<td>GTGTTGACATAAATACCA CGGGCTGTATACGA</td>
</tr>
<tr>
<td>rec A</td>
<td>CACTGATACTGATGAA GCATACAGTATAATTG</td>
</tr>
<tr>
<td>tac</td>
<td>CTGTTGACAATTAATCAT CGGCTGTATAAAGT</td>
</tr>
<tr>
<td>tacII</td>
<td>CTGTTGACATTAATCAT CGAACATGTAAATGT</td>
</tr>
</tbody>
</table>

**Fig. 5.6** The base sequence of the −10 and −35 regions of four natural promoters, two hybrid promoters and the consensus promoter.
this consensus sequence consists of the −35 region (5'-TTGACA-3') and the −10 region, or Pribnow box (5'-TATAAT). RNA polymerase must bind to both sequences to initiate transcription. The strength of a promoter, i.e. how many RNA copies are synthesized per unit time per enzyme molecule, depends on how close its sequence is to the consensus. While the −35 and −10 regions are the sites of nearly all mutations affecting promoter strength, other bases flanking these regions can affect promoter activity (Hawley & McClure 1983, Dueschle et al. 1986, Kelty & Rosenberg 1987). The distance between the −35 and −10 regions is also important. In all cases examined, the promoter was weaker when the spacing was increased or decreased from 17 bp.

Upstream (UP) elements located 5' of the −35 hexamer in certain bacterial promoters are A+T-rich sequences that increase transcription by interacting with the α subunit of RNA polymerase. Gourse et al. (1998) have identified UP sequences conferring increased activity to the rrrn core promoter. The best UP sequence was portable and increased heterologous protein expression from the lac promoter by a factor of 100.

Once RNA polymerase has initiated transcription at a promoter, it will polymerize ribonucleotides until it encounters a transcription-termination site in the DNA. Bacterial DNA has two types of transcription-termination site: factor-independent and factor-dependent. As their names imply, these types are distinguished by whether they work with RNA polymerase and DNA alone or need other factors before they can terminate transcription. The factor-independent transcription terminators are easy to recognize because they have similar sequences: an inverted repeat followed by a string of A residues (Fig. 5.7). Transcription is terminated in the string of A residues, resulting in a string of U residues at the 3' end of the mRNA. The factor-dependent transcription terminators have very little sequence in common with each other. Rather, termination involves interaction with one of the three known E. coli termination factors, Rho (ρ), Tau (τ) and NusA. Most expression vectors incorporate a factor-independent termination sequence downstream from the site of insertion of the cloned gene.

**Vectors for making RNA probes**

Although single-stranded DNA can be used as a sequence probe in hybridization experiments, RNA probes are preferred. The reasons for this are that the rate of hybridization and the stability are far greater for RNA–DNA hybrids compared with DNA–DNA hybrids. To make an RNA probe, the relevant gene sequence is cloned in a plasmid vector such that it is under the control of a phage promoter. After purification, the plasmid is linearized with a suitable restriction enzyme and then incubated with the phage RNA polymerase and the four ribonucleoside triphosphates (Fig. 5.8). No transcription terminator is required because the RNA polymerase will fall off the end of the linearized plasmid.

There are three reasons for using a phage promoter. First, such promoters are very strong, enabling large amounts of RNA to be made *in vitro*. Secondly, the phage promoter is not recognized by the *E. coli* RNA polymerase and so no transcription will occur inside the cell. This minimizes any selection of variant inserts. Thirdly, the RNA polymerases encoded by phages such as SP6, T7 and T3 are much simpler molecules to handle than the *E. coli* enzyme, since the active enzyme is a single polypeptide.

If it is planned to probe RNA or single-stranded DNA sequences, then it is essential to prepare RNA probes corresponding to both strands of the insert. One way of doing this is to have two different clones corresponding to the two orientations of the insert. An alternative method is to use a cloning vector in which the insert is placed between two different, opposing phage promoters (e.g. T7/T3 or T7/SP6) that flank a multiple cloning sequence (see Fig. 5.5). Since each of the two promoters is recognized by a different RNA polymerase, the direction of transcription is determined by which polymerase is used.

![Fig. 5.7](image-url) Structure of a factor-independent transcriptional terminator.
Fig. 5.8 Method for preparing RNA probes from a cloned DNA molecule using a phage SP6 promoter and SP6 RNA polymerase.

Fig. 5.9 Structure and use of the LITMUS vectors for making RNA probes. (a) Structure of the LITMUS vectors showing the orientation and restriction sites of the four polylinkers.
A further improvement has been introduced by Evans et al. (1995). In their LITMUS vectors, the polylinker regions are flanked by two modified T7 RNA polymerase promoters. Each contains a unique restriction site (SpeI or AflII) that has been engineered into the T7 promoter consensus sequence such that cleavage with the corresponding endonuclease inactivates that promoter. Both promoters are active despite the presence of engineered sites. Selective unidirectional transcription is achieved by simply inactivating the other promoter by digestion with SpeI or AflII prior to in vitro transcription (Fig. 5.9). Since efficient labelling of RNA probes demands that the template be linearized prior to transcription, at a site downstream from the insert, cutting at the site within the undesired promoter performs both functions in one step. Should the cloned insert contain either an SpeI or an AflII site, the unwanted promoter can be inactivated by cutting at one of the unique sites within the poly linker.

**Vectors for maximizing protein synthesis**

Provided that a cloned gene is preceded by a promoter recognized by the host cell, then there is a high probability that there will be detectable synthesis of the cloned gene product. However, much of the interest in the application of recombinant DNA technology is the possibility of facile synthesis of large quantities of protein, either to study its properties or because it has commercial value. In such instances, detectable synthesis is not sufficient; rather, it must be maximized. The factors affecting the level of expression of a cloned gene are shown in Table 5.2 and are reviewed by Baneyx (1999). Of these factors, only promoter strength is considered here.

### Table 5.2 Factors affecting the expression of cloned genes.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter strength</td>
<td>Page 74</td>
</tr>
<tr>
<td>Transcriptional termination</td>
<td>Page 71</td>
</tr>
<tr>
<td>Plasmid copy number</td>
<td>Page 45, Chapter 4</td>
</tr>
<tr>
<td>Plasmid stability</td>
<td>Page 47, Chapter 4</td>
</tr>
<tr>
<td>Host-cell physiology</td>
<td>Chapters 4 &amp; 5</td>
</tr>
<tr>
<td>Translational initiation sequences</td>
<td>Box 5.1, page 77</td>
</tr>
<tr>
<td>Codon choice</td>
<td>Box 5.1, page 77</td>
</tr>
<tr>
<td>mRNA structure</td>
<td>Box 5.1, page 77</td>
</tr>
</tbody>
</table>
When maximizing gene expression it is not enough to select the strongest promoter possible: the effects of overexpression on the host cell also need to be considered. Many gene products can be toxic to the host cell even when synthesized in small amounts. Examples include surface structural proteins (Beck & Bremer 1980), proteins, such as the PolA gene product, that regulate basic cellular metabolism (Murray & Kelley 1979), the cystic fibrosis transmembrane conductance regulator (Gregory et al. 1990) and lentivirus envelope sequences (Cunningham et al. 1993). If such cloned genes are allowed to be expressed there will be a rapid selection for mutants that no longer synthesize the toxic protein. Even when overexpression of a protein is not toxic to the host cell, high-level synthesis exerts a metabolic drain on the cell. This leads to slower growth and hence in culture there is selection for variants with lower or no expression of the cloned gene because these will grow faster. To minimize the problems associated with high-level expression, it is usual to use a vector in which the cloned gene is under the control of a regulated promoter.

Many different vectors have been constructed for regulated expression of gene inserts but most of those in current use contain one of the following controllable promoters: \( \lambda P_L \), T7, trc (tac) or BAD. Table 5.3 shows the different levels of expression that can be achieved when the gene for chloramphenicol acetyltransferase (CAT) is placed under the control of these three promoters.

The trc and tac promoters are hybrid promoters derived from the lac and trp promoters (Brosius 1984). They are stronger than either of the two parental promoters because their sequences are more like the consensus sequence. Like lac, the trc and tac promoters are inducible by lactose and isopropyl-\( \beta \)-d-thiogalactoside (IPTG).

<table>
<thead>
<tr>
<th>Promoter</th>
<th>Uninduced level of CAT</th>
<th>Induced level of CAT</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda P_L )</td>
<td>0.0275</td>
<td>28.18</td>
<td>1025</td>
</tr>
<tr>
<td>trc</td>
<td>1.10</td>
<td>5.15</td>
<td>4.7</td>
</tr>
<tr>
<td>T7</td>
<td>1.14</td>
<td>15.40</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Table 5.3 Control of expression of chloramphenicol acetyltransferase (CAT) in *E. coli* by three different promoters. The levels of CAT are expressed as \( \mu g/\text{mg} \) total protein.

When selecting a promoter, it is important to consider not only the level of expression but also the effects of any toxic products. The choice of vector and promoter is critical in ensuring that the gene is expressed in a controlled manner. The table above provides a useful guide to the different levels of expression that can be achieved with different promoters. However, the actual level of expression will depend on the specific gene and the host cell being used.

The \( \lambda P_L \) promoter system combines very tight transcriptional control with high levels of gene expression. This is achieved by putting the cloned gene under the control of the \( P_L \) promoter carried on a vector, while the \( P_L \) promoter is controlled by a \( cI \) repressor gene in the *E. coli* host. This \( cI \) gene is itself under the control of the tryptophan (trp) promoter (Fig. 5.11). In the absence of exogenous tryptophan, the \( cI \) gene is transcribed and the \( cI \) repressor binds to the \( P_L \) promoter, preventing expression of the gene. However, in the presence of tryptophan, the \( cI \) gene is repressed and the gene is expressed.

The pET vectors are a family of expression vectors that utilize phage T7 promoters to regulate synthesis of cloned gene products (Studier et al. 1990). The general strategy for using a pET vector is shown in Fig. 5.10. To provide a source of phage-T7 RNA polymerase, *E. coli* strains that contain gene 1 of the phage have been constructed. This gene is cloned downstream of the lac promoter, in the chromosome, so that the phage polymerase will only be synthesized following IPTG induction. The newly synthesized T7 RNA polymerase will then transcribe the foreign gene in the pET plasmid. If the protein product of the cloned gene is toxic, it is possible to minimize the uninduced level of T7 RNA polymerase. First, a plasmid compatible with pET vectors is selected and the T7 lysS gene is cloned in it. When introduced into a host cell carrying a pET plasmid, the lysS gene will bind any residual T7 RNA polymerase (Studier 1991, Zhang & Studier 1997). Also, if a lac operator is placed between the T7 promoter and the cloned gene, this will further reduce transcription of the insert in the absence of IPTG (Dubendorff & Studier 1991). Improvements in the yield of heterologous proteins can sometimes be achieved by use of selected host cells (Miroux & Walker 1996).
cloned gene. Upon addition of tryptophan, the \textit{trp} repressor binds to the \textit{cl} gene, preventing synthesis of the \textit{cl} repressor. In the absence of \textit{cl} repressor, there is a high level of expression from the very strong \textit{P}_{L} promoter.

The \textit{pBAD} vectors, like the ones based on \textit{P}_{L} promoter, offer extremely tight control of expression of cloned genes (Guzman et al. 1995). The \textit{pBAD} vectors carry the promoter of the \textit{araBAD} (arabinose) operon and the gene encoding the positive and
negative regulator of this promoter, araC. AraC is a transcriptional regulator that forms a complex with L-arabinose. In the absence of arabinose, AraC binds to the O₂ and I₁ half-sites of the araBAD operon, forming a 210 bp DNA loop and thereby blocking transcription (Fig. 5.12). As arabinose is added to the growth medium, it binds to AraC, thereby releasing the O₂ site. This in turn causes AraC to bind to the I₂ site adjacent to the I₁ site. This releases the DNA loop and allows transcription to begin. Binding of AraC to I₁ and I₂ is activated in the presence of cAMP activator protein (CAP) plus cAMP. If glucose is added to the growth medium, this will lead to a repression of cAMP synthesis, thereby decreasing expression from the araBAD promoter. Thus one can titrate the level of cloned gene product by varying the glucose and arabinose content of the growth medium (Fig. 5.12). According to Guzman et al. (1995), the pBAD vectors permit fine-tuning of gene expression. All that is required is to change the sugar composition of the medium. However, this is disputed by others (Siegele & Hu 1997, Hashemzadeh-Bonehi et al. 1998).

Many of the vectors designed for high-level expression also contain translation-initiation signals optimized for *E. coli* expression (see Box 5.1).

**Vectors to facilitate protein purification**

Many cloning vectors have been engineered so that the protein being expressed will be fused to another protein, called a *tag*, which can be used to facilitate protein purification. Examples of tags include glutathione-S-transferase, the MalE (maltose-binding) protein and multiple histidine residues, which can easily be purified by affinity chromatography. The tag vectors are usually constructed so that the coding sequence for an amino acid sequence cleaved by a specific protease is inserted between the coding sequence for the tag and the gene being expressed. After purification, the tag protein can be cleaved off with the specific protease to leave a normal or nearly normal protein. It is also possible to include in the tag a protein sequence that can be assayed easily. This permits assay of the cloned gene product when its activity is not known or when the usual assay is inconvenient. Three different examples of tags are given below. The reader requiring a more detailed insight should consult the review by LaVallie and McCoy (1995).

To use a polyhistidine fusion for purification, the gene of interest is first engineered into a vector in which there is a polylinker downstream of six histidine residues and a proteolytic cleavage site. In the example shown in Fig. 5.13, the cleavage site is that for enterokinase. After induction of synthesis of the fusion protein, the cells are lysed and the viscosity of the lysate is reduced by nuclease treatment. The lysate is then applied to a column containing immobilized divalent nickel, which selectively binds the polyhistidine tag. After washing away any contaminating proteins, the fusion protein is eluted from the column and treated with enterokinase to release the cloned gene product.

![Fig. 5.12 Regulation of the pBAD promoter. (a) The conformational changes that take place on addition of arabinose. (b) Western blot showing the increase in synthesis of a cloned gene product when different levels of arabinose are added to a culture of the host cell.](image)
High-level expression of a cloned gene requires more than a strong promoter. The mRNA produced during transcription needs to be effectively translated into protein. Although many factors can influence the rate of translation, the most important is the interaction of the ribosome with the bases immediately upstream from the initiation codon of the gene. In bacteria, a key sequence is the ribosome-binding site or Shine–Dalgarno (S-D) sequence. The degree of complementarity of this sequence with the 16S rRNA can affect the rate of translation (De Boer & Hui 1990). Maximum complementarity occurs with the sequence 5′-UAAGGAGG-3′ (Ringquist et al. 1992). The spacing between the S-D sequence and the initiation codon is also important. Usually there are five to 10 bases, with eight being optimal. Decreasing the distance below 4 bp or increasing it beyond 14 bp can reduce translation by several orders of magnitude.

Translation is affected by the sequence of bases that follow the S-D site (De Boer et al. 1983b). The presence of four A residues or four T residues in this position gave the highest translational efficiency. Translational efficiency was 50% or 25% of maximum when the region contained, respectively, four C residues or four G residues.

The composition of the triplet immediately preceding the AUG start codon also affects the efficiency of translation. For translation of β-galactosidase mRNA, the most favourable combinations of bases in this triplet are UAU and CUU. If UUC, UCA or AGG replaced UAU or CUU, the level of expression was 20-fold less (Hui et al. 1984).

The codon composition following the AUG start codon can also affect the rate of translation. For example, a change in the third position of the fourth codon of a human γ-interferon gene resulted in a 30-fold change in the level of expression (De Boer & Hui 1990). Also, there is a strong bias in the second codon of many natural mRNAs, which is quite different from the general bias in codon usage. Highly expressed genes have AAA (Lys) or GCU (Ala) as the second codon. Devlin et al. (1988) changed all the G and C nucleotides for the first four codons of a granulocyte colony-stimulating factor gene and expression increased from undetectable to 17% of total cell protein.

Sequences upstream from the S-D site can affect the efficiency of translation of certain genes. In the E. coli md gene there is a run of eight uracil residues. Changing two to five of these residues has no effect on mRNA levels but reduces translation by up to 95% (Zhang & Deutscher 1992). Etchegaray and Inouye (1999) have identified an element downstream of the initiation codon, the downstream box, which facilitates formation of the translation-initiation complex. The sequence of the 3′ untranslated region of the mRNA can also be important. If this region is complementary to sequences within the gene, hairpin loops can form and hinder ribosome movement along the messenger.

The genetic code is degenerate, and hence for most of the amino acids there is more than one codon. However, in all genes, whatever their origin, the selection of synonymous codons is distinctly non-random (for reviews, see Kurland 1987, Ernst 1988 and McPherson 1988). The bias in codon usage has two components: correlation with tRNA availability in the cell, and non-random choices between pyrimidine-ending codons. Ikemura (1981a) measured the relative abundance of the 26 known tRNAs of E. coli and found a strong positive correlation between tRNA abundance and codon choice. Later, Ikemura (1981b) noted that the most highly expressed genes in E. coli contain mostly those codons corresponding to major tRNAs but few codons of minor tRNAs. In contrast, genes that are expressed less well use more suboptimal codons. Forman et al. (1998) noted significant misincorporation of lysine, in place of arginine, when the rare AGA codon was included in a gene overexpressed in E. coli. It should be noted that the bias in codon usage even extends to the stop codons (Sharp & Bulmer 1988). UAA is favoured in genes expressed at high levels, whereas UAG and UGA are used more frequently in genes expressed at a lower level.

For a review of translation the reader should consult Kozak (1999).
For the cloned gene to be expressed correctly, it has to be in the correct translational reading frame. This is achieved by having three different vectors, each with a polylinker in a different reading frame (see Fig. 5.13). Enterokinase recognizes the sequence (Asp)_4Lys and cleaves immediately after the lysine residue. Therefore, after enterokinase cleavage, the cloned gene protein will have a few extra amino acids at the N terminus. If desired, the cleavage site and polyhistidines can be synthesized at the C terminus. If the cloned gene product itself contains an enterokinase cleavage site, then an alternative protease, such as thrombin or factor Xa, with a different cleavage site can be used.

To facilitate assay of the fusion proteins, short antibody recognition sequences can be incorporated into the tag between the affinity label and the protease cleavage site. Some examples of recognizable sequences include:

- **Frame-dependent variations**

<table>
<thead>
<tr>
<th>Frame</th>
<th>Site</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>pBAD/His</td>
<td>A,B,C</td>
<td>4.1 kb</td>
</tr>
<tr>
<td>2</td>
<td>ColE1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ampicillin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>araC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>PBAD ATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6xHis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Epitope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>MCS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 5.13** Structure of a vector (pBAD/His, In Vitrogen Corporation) designed for the expression of a cloned gene as a fusion protein containing a polyhistidine sequence. Three different variants (A, B, C) allow the insert to be placed in each of the three translational reading frames. The sequence shaded pink shows the base sequence which is altered in each of the three vectors. The lightly-shaded box (AGATCT) is the Bgl II site of the polylinker. Note that the initial A residue of the restriction site is at a different point in the triplet codon in each of the three sequences.
epitopes are given in Table 5.4. These antibodies can be used to detect, by western blotting, fusion proteins carrying the appropriate epitope. Note that a polyhistidine tag at the C terminus can function for both assay and purification.

Biotin is an essential cofactor for a number of carboxylases important in cell metabolism. The biotin in these enzyme complexes is covalently attached at a specific lysine residue of the biotin carboxylase carrier protein. Fusions made to a segment of the carrier protein are recognized in *E. coli* by biotin ligase, the product of the *birA* gene, and biotin is covalently attached in an ATP-dependent reaction. The expressed fusion protein can be purified using streptavidin affinity chromatography (Fig. 5.14). *E. coli* expresses a single endogenous biotinylated protein, but it does not bind to streptavidin in its native configuration, making the affinity purification highly specific for the recombinant fusion protein. The presence of biotin on the fusion protein has an additional advantage: its presence can be detected with enzymes coupled to streptavidin.

The affinity purification systems described above suffer from the disadvantage that a protease is required to separate the target protein from the affinity tag. Also, the protease has to be separated from the protein of interest. Chong *et al.* (1997, 1998) have described a unique purification system that has neither of these disadvantages. The system utilizes a protein splicing element, an intein, from the *Saccharomyces cerevisiae* VMA1 gene (see Box 5.2). The intein is modified such that it undergoes a self-cleavage reaction at its N terminus at low temperatures in the presence of thiols, such as cysteine, dithiothreitol or β-mercaptoethanol. The gene encoding the target protein is inserted into a multiple cloning site (MCS) of a vector to create a fusion between the C terminus of the target gene and the N terminus of the gene encoding the intein. DNA encoding a small (5 kDa) chitin-binding domain from *Bacillus circulans* was added to the C terminus of the intein for affinity purification (Fig. 5.15).

The above construct is placed under the control of an IPTG-inducible T7 promoter. When crude extracts from induced cells are passed through a chitin column, the fusion protein binds and all contaminating proteins are washed through. The fusion is then induced to undergo intein-mediated self-cleavage on the column by incubation with a thiol. This releases the target protein, while the intein chitin-binding domain remains bound to the column.

**Vectors to promote solubilization of expressed proteins**

One of the problems associated with the overproduction of proteins in *E. coli* is the sequestration of the product into insoluble aggregates or ‘inclusion bodies’ (Fig. 5.16). They were first reported in strains overproducing insulin A and B chains (Williams *et al.* 1982). At first, their formation was thought to be restricted to the overexpression of heterologous proteins in *E. coli*, but they can form in the presence of high levels of normal *E. coli* proteins, e.g. subunits of RNA polymerase (Gribskov & Burgess 1983). Two parameters that can be manipulated to reduce inclusion-body formation are temperature and growth rate. There are a number of reports which show that lowering the temperature of growth increases the yield of correctly folded, soluble protein (Schein & Noteborn 1988, Takagi *et al.* 1988, Schein 1991). Media compositions and pH values that reduce the growth rate also reduce inclusion-body formation. Renaturation of misfolded proteins can sometimes be achieved following solubilization in guanidinium hydrochloride (Lilie *et al.* 1998).
Fig. 5.14  Purification of a cloned gene product synthesized as a fusion to the biotin carboxylase carrier protein (tag). See text for details.
Three ‘genetic’ methods of preventing inclusion-body formation have been described. In the first of these, the host cell is engineered to overproduce a chaperon (e.g. DnaK, GroEL or GroES proteins) in addition to the protein of interest (Van Dyk et al. 1989, Blum et al. 1992, Thomas et al. 1997). Castanie et al. (1997) have developed a series of vectors which are compatible with pBR322-type plasmids and which encode the overproduction of chaperons. These vectors can be used to test the effect of chaperons on the solubilization of heterologous gene products. Even with excess chaperon there is no guarantee of proper folding. The second method involves making minor changes to the amino acid sequence of the target protein. For example, cysteine-to-serine changes in fibroblast growth factor minimized inclusion-body formation (Rinas et al. 1992). The third method is derived from the observation that many proteins produced as insoluble aggregates in their native state are synthesized in soluble form as thioredoxin fusion proteins (LaVallie et al. 1993). More recently, Davis et al. (1999) have shown that the NusA and GrpE proteins, as well as bacterioferritin, are even better than thioredoxin at solubilizing proteins expressed at a high level. Kapust and Waugh (1999) have reported that the maltose-binding protein is also much better than thioredoxin.

Building on the work of LaVallie et al. (1993), a series of vectors has been developed in which the gene of interest is cloned into an MCS and the gene product is produced as a thioredoxin fusion protein with an enterokinase cleavage site at the fusion point. After synthesis, the fusion protein is released from the producing cells by osmotic shock and purified. The desired protein is then released by enterokinase cleavage. To simplify the purification of thioredoxin fusion proteins, Lu et al. (1996)
systematically mutated a cluster of surface amino acid residues. Residues 30 and 62 were converted to histidine and the modified ('histidine patch') thioredoxin could now be purified by affinity chromatography on immobilized divalent nickel. An alternative purification method was developed by Smith et al. (1998). They synthesized a gene in which a short biotinylation peptide is fused to the N terminus of the thioredoxin gene to generate a new protein called BIOTRX. They constructed a vector carrying the BIOTRX gene, with an MCS at the C terminus, and the birA gene. After cloning a gene in the MCS, a fused protein is produced which can be purified by affinity chromatography on streptavidin columns.

An alternative way of keeping recombinant proteins soluble is to export them to the periplasmic space (see next section). However, even here they may still be insoluble. Barth et al. (2000) solved this problem by growing the producing bacteria under osmotic stress (4% NaCl plus 0.5 mol/l sorbitol) in the presence of compatible solutes. Compatible solutes are low-molecular-weight osmolytes, such as glycine betaine, that occur naturally in halophilic bacteria and are known to protect proteins at high salt concentrations. Adding glycine betaine for the cultivation of *E. coli* under osmotic stress not only allowed the bacteria to grow under these otherwise inhibitory conditions but also produced a periplasmic environment for the generation of correctly folded recombinant proteins.

**Vectors to promote protein export**

Gram-negative bacteria such as *E. coli* have a complex wall–membrane structure comprising an inner, cytoplasmic membrane separated from an outer membrane by a cell wall and periplasmic space.
Secreted proteins may be released into the periplasm or integrated into or transported across the outer membrane. In *E. coli* it has been established that protein export through the inner membrane to the periplasm or to the outer membrane is achieved by a universal mechanism known as the general export pathway (GEP). This involves the *sec* gene products (for review see Lory 1998). Proteins that enter the GEP are synthesized in the cytoplasm with a signal sequence at the N terminus. This sequence is cleaved by a signal or leader peptidase during transport. A signal sequence has three domains: a positively charged amino-terminal region, a hydrophobic core, consisting of five to 15 hydrophobic amino acids, and a leader peptidase cleavage site. A signal sequence attached to a normally cytoplasmic protein will direct it to the export pathway.

Many signal sequences derived from naturally occurring secretory proteins (e.g. OmpA, OmpF, PeB, β-lactamase, alkaline phosphatase and phage M13 gIII) support the efficient translocation of heterologous peptides across the inner membrane when fused to their amino termini. In some cases, however, the preproteins are not readily exported and either become ‘jammed’ in the inner membrane, accumulate in precursor inclusion bodies or are rapidly degraded within the cytoplasm. In practice, it may be necessary to try several signal sequences (Berges *et al.* 1996) and/or overproduce different chaperons to optimize the translocation of a particular heterologous protein. A first step would be to try the secretion vectors offered by a number of molecular-biology suppliers and which are variants of the vectors described above.

**Fig. 5.16** Inclusions of Trp polypeptide–proinsulin fusion protein in *E. coli*. (Left) Scanning electron micrograph of cells fixed in the late logarithmic phase of growth; the inset shows normal *E. coli* cells. (Right) Thin section through *E. coli* cells producing Trp polypeptide–insulin A chain fusion protein. (Photographs reproduced from *Science* courtesy of Dr D.C. Williams (Eli Lilly & Co.) and the American Association for the Advancement of Science.)
It is possible to engineer proteins such that they are transported through the outer membrane and are secreted into the growth medium. This is achieved by making use of the type I, Sec-independent secretion system. The prototype type I system is the haemolysin transport system, which requires a short carboxy-terminal secretion signal, two translocators (HlyB and D), and the outer-membrane protein TolC. If the protein of interest is fused to the carboxy-terminal secretion signal of a type I-secreted protein, it will be secreted into the medium provided HlyB, HlyD and TolC are expressed as well (Spreng et al. 2000). An alternative presentation of recombinant proteins is to express them on the surface of the bacterial cell using any one of a number of carrier proteins (for review, see Cornelis 2000).

Putting it all together: vectors with combinations of features

Many of the vectors in current use, particularly those that are commercially available, have combinations of the features described in previous sections. Two examples are described here to show the connection between the different features. The first example is the LITMUS vectors that were described earlier (p. 73) and which are used for the generation of RNA probes. They exhibit the following features:

- The polylinkers are located in the lacZα gene and inserts in the polylinker prevent α-complementation. Thus blue/white screening (see Box 3.2 on p. 35) can be used to distinguish clones with inserts from those containing vector only.
- The LITMUS polylinkers contain a total of 32 unique restriction sites. Twenty-nine of these enzymes leave four-base overhangs and three leave blunt ends. The three blunt cutting enzymes have been placed at either end of the polylinker and in the middle of it.
- The vectors carry both the pUC and the M13 ori regions. Under normal conditions the vector replicates as a double-stranded plasmid but, on infection with helper phage (M1 3K07), single-stranded molecules are produced and packaged in phage protein.
- The single-stranded molecules produced on helper-phage addition have all the features necessary for DNA sequencing (see p. 123).

- The vectors are small (< 3 kb) and with a pUC ori have a high copy number.

The second example is the PinPoint series of expression vectors (Fig. 5.17). These vectors have the following features:

- Expression is under the control of either the T7 or the tac promoter, allowing the user great flexibility of control over the synthesis of the cloned gene product.
- Some of them carry a DNA sequence specifying synthesis of a signal peptide.
- Presence of an MCS adjacent to a factor-Xa cleavage site.
- Synthesis of an N-terminal biotinylated sequence to facilitate purification.
- Three different vectors of each type permitting translation of the cloned gene insert in each of the three reading frames.
- Presence of a phage SP6 promoter distal to the MCS to permit the synthesis of RNA probes complementary to the cloned gene. Note that the orientation of the cloned gene is known and so the RNA probe need only be synthesized from one strand.

What is absent from these vectors is an M13 origin of replication to facilitate synthesis of single strands for DNA sequencing.